An evaluation of the role of statistical measures and frequency for MWE identification

Sandra Antunes and Amália Mendes

Centre for Linguistics at the University of Lisbon, Portugal

{sandra.antunes, amalia.mendes}@clul.ul.pt

Abstract

We report on an experiment to evaluate the role of statistical association measures and frequency for the identification of MWE. We base our evaluation on a lexicon of 14.000 MWE comprising different types of word combinations: collocations, nominal compounds, light verbs + predicate, idioms, etc. These MWE were manually validated from a list of n-grams extracted from a 50 million word corpus of Portuguese (a subcorpus of the Reference Corpus of Contemporary Portuguese), using several criteria: syntactic fixedness, idiomaticity, frequency and Mutual Information measure, although no threshold was established, either in terms of group frequency or MI. We report on MWE that were selected on the basis of their syntactic and semantics properties while the MI or both the MI and the frequency show low values, which would constitute difficult cases to establish a cutting point. We analyze the MI values of the MWE selected in our gold dataset and, for some specific cases, compare these values with two other statistical measures.

1. Introduction

Many studies and theories regarding the phenomenon of multiword expressions (MWE) have been pursued since Firth's well-known "you shall know a word by the company it keeps" (Firth, 1957:11). Sinclair (1991:110) strengthens this idea pointing out that words frequently and systematically attract each other, creating complex patterns of associations and making meanings by their combinations, which results in pre-constructed phrases that speakers frequently use in their conversations (idiom principle).

We will use the term MWE as including different types of word combinations (collocations, nominal compound, light verbs, idioms, etc.) that may present certain properties, such as lexical and syntactic fixedness (which can be observed through the possibility of replacing elements, inserting modifiers, changing the syntagmatic structure, etc.), total or partial loss of compositional meaning and frequency of occurrence (which may reveal sets of favoured co-occurring forms, showing that they may be in their way to a possible fixedness).

It is now widely known that MWE play a crucial role in language and that great part of a speaker's lexicon is composed by these word associations (Jackendoff, 1997; Fellbaum, 1998). Their analysis has been carried out in several areas, ranging from psycholinguistics, second language teaching, lexicography or computational linguistics. But for linguistic research to be successful, one of the questions to be answered is how to determine the significant word combinations of a language that are worthy of analysis. Nowadays, the availability of large amounts of data and the development of corpus-based approaches make it possible to use statistical methods (such as Mutual Information (MI), Log-Likelihood, Chi-Square (χ^2), T-test or Permutation Entropy (PE)) to automatically identify MWE and to measure how closely related the words are. However, given the statistical results and the human empirical knowledge, another question arises: how well do these statistical measures perform in identifying significant MWE and distinguishing them from non-MWE; and what criteria should be used when the numbers do not cover expressions that one might think are indeed significant?

Based on a lexical dataset of MWE manually selected (Mendes et al., 2006), our goal is to discuss some difficult cases related to MI and frequency values when applied to the selection of significant n-grams.

The paper is structured as follows: we first review some experiments in the evaluation of different association measures in section 2, we then describe the corpus and the methodology adopted for compiling the lexical dataset of MWE in section 3. In section 4, we discuss the MI values of the set of manually selected MWE and, for some specific cases, we compare the MI values with two other statistical measures (section 5).

2. Related work

There are several approaches taken by researchers regarding the extraction of MWE from textual data. Dunning (1993) briefly refers three categories: (i) the collection of large amounts of text in order to make statistical measures perform well; (ii) the application of a simple statistical analysis on relatively small amounts of text and the empirical correction of errors (but statistical measures can overestimate the significance of some events when the counts involved are small); (iii) no use of statistical analysis.

Regarding statistical analysis of text, many methods have been used. Several studies have evaluated and compared different methods of automatic extraction of MWE (Dunning, 1993; Evert & Krenn, 2001; Pearce, 2002; Villavicencio et al., 2007). However, as Villavicencio et al. (2007:1034) point out, "given the heterogeneousness of the different phenomena that are considered to be MWEs, there is no consensus about which method is best suited for which type of MWE, and if there is a single method that can be successfully use for

any kind of MWE". These authors evaluated the application of three different statistical measures (MI, χ^2 and PE) for automatically identifying MWE from four different corpora: two web generated corpora (using Google and Yahoo) and a sample of the BNC[1] (a more homogenous and balanced corpus). The results were that: (i) the different measures sorted the expressions very differently; (ii) only MI and PE seem to differentiate between MWE and non-MWE; (iii) a larger corpus may provide better samples of language use.

Evert & Krenn (2001) performed as well a comparison of several lexical association measures that were also applied to two different data sets: (i) ADJ N pairs extracted from a 816.203 word corpus; (ii) PREP N V triples extracted from a 8 million word corpus. The authors concluded that the ranking of the association measures differed depending on the type of MWE and the frequency of a word in the data. This comparison also questions the strength of Log-Likelihood in handling low-frequency data, showing indeed that none of the association measures worked well with small amounts of text.

A comparison study for Portuguese was carried out by Baptista et al. (2012). The authors used a list of fixed expressions with idiomatic meaning. The expressions had the following syntactic constitution: (i) *N0 V Prep C1* (where *N0* stands for a free subject and *C1* represents a prepositional complement with one or more words, like *ir para o galheiro* 'to ruin'; *chegar a bom porto* 'to succeed'); (ii) *N0 V Prep C2* (where *C2* represents a complex nominal, like *ir para a quinta das tabuletas* 'to die'). The authors tried to evaluate the use of T-test, χ^2 and MI for automatically identifying MWE from a 189M word newspaper corpus. However, the authors noted that approximately only half of the expressions of their list occur in the corpus (which probably results from the specific type of the expressions) and that that fact will hamper their identification based on statistical measures. Regarding the matching cases, the authors conclude that χ^2 presents better results than both T-test (which is not suitable for small data) and MI (which may be efficient regarding collocations, but is not appropriated for fixed expressions).

In conclusion, the evaluation studies point to the fact that the size and the diversity of the data seem to influence the statistical measures. Also, since practically all the comparison analysis show that different measures give different results, some authors concluded not only that "it is not possible to recommend 'the best general association measure' for ranking collocation candidates" (Pecina, 2008:57), but also that "the individual performances of these measures may well be improved if they are combined together, offering different insights into the problem" (Ramisch et al. 2008:53).

3. COMBINA-PT: corpus and MWE's extraction

[1] http://www.natcorp.ox.ac.uk/

COMBINA-PT [2] is a lexicon of 14.000 MWE semi-automatically extracted from a balanced 50 million word written corpus and manually validated. This corpus was designed as a balance subset of the Reference Corpus of Contemporary Portuguese[3] (Généreux et al., 2012), and its design and size are presented in Table 1.

CORPUS CONSTITUTION	
Newspapers	**30.000.000**
Books	**10.917.889**
Magazines	**7.500.000**
Miscellaneous	**1.851.828**
Leaflets	**104.889**
Supreme court verdicts	**313.962**
Parliament sessions	**277.586**
TOTAL	**50.966.154**

Table 1: Corpus size and design

The MWE are organized under canonical forms, and variations of these canonical forms (either lexical, syntactic or inflectional) are recorded. In total, the lexicon contains 14.153 canonical forms and 48.154 MWE variations. For each of those several examples are collected from the corpus. As described in Mendes et al. (2006), n-grams of 2, 3, 4 and 5 tokens were extracted from the corpus and statistically sorted using MI as lexical association measure (Church & Hanks, 1990). The choice of the MI measure relied on the fact that it is reported to differentiate between MWE and non-MWE (see, for instance, Villavicencio et al. (2007)). The extraction process included discontinuous 2-grams, separated by a maximum of 3 tokens, and contiguous 3 to 5-grams. Several cut-off options were also implemented when extracting the candidate set of MWE, for instance: the elimination of groups with internal punctuation and the elimination of 2-grams beginning or ending with a grammatical word. The corpus was not previously tagged with POS information nor was it lemmatized. MI was applied to word forms since MWE frequently show a preference for one of the inflected forms of the lemma.

4. The role of MI and frequency for manual validation

Previous experiments that we conducted in the automatic extraction and evaluation of MWE (Bacelar do Nascimento, 2000; Pereira & Mendes, 2002) over different Portuguese corpora showed that there was a higher concentration of good candidates around medium MI values (7-12) and that MI seems to promote very infrequent word combinations. As a starting point, we selected a list of nodes that occurred in n-grams with MI values between 8 and 10. We then manually inspected each n-gram that included one of these nodes. The validation of candidates was based on several criteria:

[2] A part of the lexicon is available at the Meta-Share repository (http://www.meta-share.eu/) under the title LEX-MWE.PT.

[3] https://www.clul.ul.pt/en/research-teams/183-reference-corpus-of-contemporary-portuguese-crpc.

syntactic (fixedness), semantic (idiomaticity) and quantificational (MI, frequency). However, we did not restrict our selection to a threshold, either in terms of group frequency or MI. Inflection and lexical-syntactic variants of the selected MWE are organized under a canonical form (e.g., *arma de fogo 'firegun'*), which is then associated with one node (e.g., *fogo* 'fire').

Our objective in this paper is to observe to what extent the quantificational data are reliable for distinguishing between MWE and non-MWE (taking non-MWE to be syntagmatic sequences which are not idiomatic, nor fixed, nor form a preferred combination of words).

MWE that responded to the syntactic and semantic criteria were frequently n-grams with medium MI values and frequency, such as examples in Table 2, in accordance to the expected behaviour of MI.

MWE	MI	Freq.
papel fundamental 'key role'	7.8	194
fonte de inspiração 'source of inspiration'	8.7	60
consequências graves 'severe consequences'	9.7	145
período homólogo 'the same period'	10.7	237
consciência tranquila 'clear consciense'	11.1	105
integridade física 'physical integrity'	12.0	136

Table 2: MWE with MI around 7-12 and high frequency

However, one of the major issues that arose during manual inspection was the fact that many n-grams that respond positively to the criteria that identify MWE show in fact an extremely low MI value (see Table 3). These are frequently sequences with one or more high-frequent words in the corpus: it is the case, for instance, of auxiliary verbs in idiomatic expressions (e.g. *estar em forma* 'to be in good shape'), support (or light) verbs followed by a predicative noun or adjective (*ter força* 'to have strength') and figurative or idiomatic uses of main verbs (*ganhar tempo* 'to save time', *ir em frente* 'go ahead'). It is also the case of nominal MWE in Table 3: the low MI value is due to the high frequency of each individual word of the expressions in the corpus. Since MI does not positively rank high frequent words, these MWE receive a low statistical significance value.

MWE	MI	Freq.
ter força 'to have strength'	2.2	306
ir em frente 'to move on'	1.4	85
ganhar tempo 'to save time'	3.1	83
estar em forma 'to be in good shape'	2.9	82
cultura geral 'general knowledge'	2.8	53
gente grande 'grown-up'	1.9	42
gente de bem 'good people'	3.3	109
lei do mais forte 'law of the jungle'	2.6	27

Table 3: MWE with low MI and high frequency

Due to this statistical property of MI, longer groups may not have higher MI values. Indeed, the presence of the high-frequent preposition *em* 'in', in the MWE *em flagrante delito* (Table 4) lowers the MI value.

However, if a non-grammatical and non-frequent word is added to the expression, the MI rises, as also shown in Table 4: both *bens de consumo corrente* and *bens de consumo duradouro* have a higher MI value than *bens de consumo*. Also, if a particular word of a MWE occurs with low isolated frequency in the corpus, it will probably bring about more striking combinations, as it happens with *fonte fidedigna* and *aborto eugénico*.

MWE	Elem.	MI	Freq.
flagrante delito 'flagrant offence'	2	15.3	85
em flagrante delito 'in flagrant offence'	3	12.9	35
bens de consumo 'consumer goods'	3	8.4	179
bens de consumo corrente 'daily consumer goods'	4	11.5	18
bens de consumo duradouro 'durable consumer goods'	4	13.8	9
fonte fidedigna 'reliable source'	2	11.1	13
aborto eugénico 'eugenics abortion'	2	14.4	20

Table 4: Comparison of MI values of MWE with 2, 3 and 4 grams

Still, in cases where the expression allows for the insertion of lexical elements (usually adverbs and quantifiers), despite the high frequency of occurrence of these items in the corpus, we also observed that the longer group may have a higher MI value (Table 5).

MWE	Elem.	MI	Freq
personalidade forte 'strong personality'	2	6.2	24
personalidade muito forte 'very strong personality'	3	8.4	5
conjunto vasto 'extensive set'	2	6.9	16
conjunto mais vasto 'more extensive set'	3	10.0	5

Table 5: longer groups with higher MI values

Looking now at cases of MWE with high MI values, Table 6 clearly shows that these values match up with MWE that include non-frequent words in the corpus and that in most cases correspond to expressions that fall within the scope of terminology register.

MWE	MI	Freq.
efluentes gasosos 'gaseous effluents'	15.1	10
cônjuge sobrevivo 'surviving spouse'	16.5	25
mucosa gástrica 'gastric mucosa'	17.6	14
fuso mitótico 'mitotic spindle'	18.4	11
organismos geneticamente modificados 'genetically modified organisms'	19.1	42
corrupção passiva para acto ilícito 'passive corruption for illicit act'	20.0	4

Table 6: MI range and frequency

Coming back to the cases of low MI in Table 3, notice that all these MWE have nevertheless high or medium frequencies. The combination of a statistical approach with raw frequency would enable us to recover these

expressions, which would otherwise be ignored as non-MWE with sole MI. But our manual validation also points to more difficult cases, when both MI and frequency have low values, as the examples in Table 7. All qualify as MWE in terms of their semantic and syntactic properties, since:

- they do not accept inflection variation in one or all of their elements,
- they restrict insertion of lexical and grammatical elements inside their structure,
- they express specific entities or qualities and their meaning is not processed compositionally,
- and they are all intuitively recognized as MWE by native speakers.

However, no such correspondence is to be found in quantitative criteria extracted from our 50 million word corpus (Table 7).

MWE	MI	Freq.
fonte de vida 'source of life'	2.7	5
de última geração 'state-of-the-art'	3.4	4
prova de fundo 'long distance race'	3.7	4
folha de serviço 'track record'	4.5	5

Table 7: MWE with low MI and low frequency

These cases pose a challenge to an automatic approach using MI and frequency for MWE selection, although quantificational information (lexical association measures and frequency) should certainly be taken into account. We already mentioned cases as the ones exemplified in Table 3 and there is no doubt that both criteria are important in cases of almost-synonym. See, for instance, the examples in Table 8: the most frequent combination has a lower MI due to the specific strength of occurrence of the word *vindouras* with the word *gerações*.

MWE	MI	Freq.
gerações seguintes 'next generations'	9.9	8
gerações futuras 'future generations'	11.3	**60**
gerações vindouras 'generations to come'	**14.9**	29
em termos gerais 'generally speaking'	8.7	**107**
em termos globais 'broadly speaking'	**9.8**	88

Table 8: Almost-synonym collocations

The observation of cases such as the ones illustrated by Table 3 and 7 led us to include, for the compilation of the COMBINA-PT lexicon, MWE with MI values and frequencies that would a priori be set aside. Observing our lexicon, the MI values range from 1.4 to 24.1, regardless the frequency of occurrence of the expression in the corpus. When we organize the selected MWE in 4 thresholds for MI values (1.0-4.9; 5.0-9.9; 10.0-14.9 and 15.0->20.0), our data corroborates the assumption that values around 5-10 concentrate the higher number of interesting MWE, as can be seen in Chart 1.

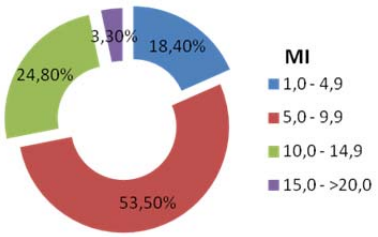

Chart1: Distribution of MWE per MI values

These values of MI account for around 50% of our gold dataset, manually selected. Almost 25% of equally valid MWE receive values between 10-15, and almost 19% values between 1 and 5. A threshold between 5-15 MI value accounts for almost 80% of our gold dataset. However, lower values still include a high number of significant MWE, proving that one can actually find significant MWE throughout all the range of values, although in different proportions (notice that Evert & Krenn (2001:190) pointed out that MI's precision remains almost constant or even increases slightly over the data).

5. Comparison with other statistical measures

Since, as far as we know, evaluation tests in the literature have been performed using different corpora, we planned to analyze some MWE from the same corpus using different statistical measures to see how the values behave within the same data. For that purpose, we compared MI with T-test and Log-Likelihood. Considering some MWE for the node *fogo* 'fire', we noticed that the values of the different statistical measures are indeed very similar: the highest and lowest rank of the three measures correspond to the same expressions, while in the middle values, the order only varies slightly (Table 9). Regarding the general arrangement of MWE, it seems that Log-Likelihood is a little closer to MI than T-test.

MWE	MI	T-test	Log-Like.
cessar fogo 'ceasefire'	13	19.8	6733.1
fogo de artifício 'firework'	12.4	11.2	1986,2
fogo cruzado 'crossfire'	**11,6**	7.4	**792.4**
arma de fogo 'firegun'	**9.6**	**9.9**	**1138.9**
debaixo de fogo 'under fire'	8.4	**8.2**	664.6
a ferro e fogo 'put to fire and sword'	7.2	7.4	449.5
linha de fogo 'firing line'	6.7	**8.1**	491.4
prova de fogo 'key test'	6.6	7.7	444.8
cor de fogo 'fire red'	6.2	5.4	199.2
abrir fogo 'to open fire'	6	4.2	121.6
mar de fogo 'sea of fire'	3.7	3.4	47.19

Table 9: MWE analyzed with MI, T-test and Log-Likelihood

Finally, coming back to the manually validated expressions where both MI and frequency have low values (Table 7), we wanted to test if the other statistical measures ranked these MWE higher, so that they could be automatically extracted. Again, the results were not very different from MI (Table 10). The main difference is the higher ranking of the expression *estar em forma* (with an auxiliary verb) by both T-test and Log-Likelihood. Also, T-test presented a good result for *cultura geral*.

MWE	MI	T-test	Log-Like.
lei do mais forte 'law of the jungle'	2.6	4.7	64.6
fonte de vida 'source of life'	2.7	5.2	85.9
cultura geral 'general knowledge'	2.8	8.6	244.2
estar em forma 'to be in good shape'	**2.9**	**15.1**	**1073.6**
de última geração 'state-of-the-art'	3.4	4.2	75.1

Table 10: MWE with low statistical values

6. Conclusion

We presented some of the issues observed during the selection of set of 14.000 MWE, extracted from a 50 million word written corpus and manually validated using MI statistical measure and frequency. Analyzing the data, it has become clear that the high/low frequency of an isolated word in the corpus would clearly influence the MI value of the group in which it occurs. But one of the major challenges was the existence of significant expressions with extremely low MI values and low frequency that would be hardly recovered automatically. We reported the distribution of the selected MWE over 4 thresholds for MI values and showed that our data corroborates our initial hypothesis that medium values (around 5-10) concentrate the higher number of interesting MWE. Furthermore, MI values between 5-15 account for almost 80% of our dataset.

An automatic selection process would have to deal with the bottleneck of correctly identifying the remaining 20% of significant MWE. Taking some examples into consideration, we compared MI values with T-test and Log-Likelihood. We didn't find major significant differences between the results, except for one case.

In the future, we plan to analyze the distribution of the MI values and to cross this information with the different types of internal structure of MWE. The same process will be performed with the two other statistical measures discussed in this paper. We believe the result of this validation work can be important for research on the automatic extraction of MWE from corpus data and can help shed some light on the importance of quantificational methods in this area.

Acknowledgements

This work was partially supported by national funds through FCT – Fundação para a Ciência e Technologia, under project PEst-OE/LIN/UI0214/2013. We would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

Bacelar do Nascimento, M.F. (2000). Exemples de combinatoires lexicales établis pour l'écrit et l'oral à Lisbonne. In Bilger, M. (ed.). *Corpus, Méthodologie et Applications Linguistiques*. Paris: H. Champion et Presses Universitaires de Perpignan 2000, pp. 237--261.

Baptista, J., Vale, O.A. & Mamede, N. (2012). Identificação de Expressões Fixas em Corpora: até onde podem ir os métodos estatísticos? In T. M. G. Shepherd et al. (eds.). *Caminhos da Linguística de Corpus*. Campinas, SP: Mercado de Letras, pp. 177--190.

Church K. & Hanks P. (1990). Word Association Norms, Mutual Information and Lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada, pp. 76--83.

Dunning, T. (1993). Accurate Models for the Statistics of Surprise and Coincidence. *Computational Linguistics, (19)1*, pp. 61--74.

Evert S. & Krenn B. (2001). Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, pp. 188--195.

Fellbaum, C. (1998). *An WordNet Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Généreux, M., Hendrickx I. & Mendes A. (2012). A Large Portuguese Corpus On-Line: Cleaning and Preprocessing. In H. Caseli et al. (eds.). *Computational Processing of the Portuguese Language. Proceedings of the 10th International Conference PROPOR2012*. Berlin, Heidelberg: Springer-Verlag, pp. 113--120.

Firth, R.J. (1957). Modes of meaning. *Papers in Linguistics 1934-1951*. London, Oxford University Press, pp. 190--215.

Jackendoff, R. (1997). *The Architecture of the Language Faculty*. The MIT Press, Cambridge, MA.

Mendes A., Antunes, S., Bacelar do Nascimento, M.F., Casteleiro, J.M., Pereira, L. & Sá, T. (2006). COMBINA-PT: A Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy, pp. 1900--1905.

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. In *Proceedings of the Third International Conference on Language Resources and Evaluation.* Las Palmas, Spain, pp.13--18.

Pecina, P. (2008). A Machine Learning Approach to Multiword Expression Extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*. Marrakech, Morocco, pp. 54--57.

Pereira L. & Mendes, A. (2002). An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications. In *Proceedings of the 10th International Congress of the European Association for Lexicography*. Copenhagen, Denmark, vol. II, pp. 841--849.

Ramisch, C., Schreiner, P., Idiart, M. & Villavicencio, A. (2008). An Evaluation of Methods for the Extraction of Multiword Expressions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*. Marrakech, Morocco, pp. 50--53.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M. & Ramisch, C. (2007). Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, pp. 1034--1043.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford, Oxford University Press.